

Opinion Conflicts: An Effective Route to Detect Incivility in Twitter

SUMAN KALYAN MAITY, Northwestern University, USA

AISHIK CHAKRABORTY, McGill University, Canada

PAWAN GOYAL, Indian Institute of Technology Kharagpur, India

ANIMESH MUKHERJEE, Indian Institute of Technology Kharagpur, India

In Twitter, there is a rising trend in abusive behavior which often leads to incivility¹. This trend is affecting users mentally and as a result they tend to leave Twitter and other such social networking sites thus depleting the active user base. In this paper, we study factors associated with incivility. We observe that the act of incivility is *highly correlated with the opinion differences between the account holder (i.e., the user writing the incivil tweet) and the target (i.e., the user for whom the incivil tweet is meant for or targeted), toward a named entity*. We introduce a character level CNN model and incorporate the *entity-specific sentiment information* for efficient incivility detection which significantly outperforms multiple baseline methods achieving an impressive accuracy of **93.3%** (4.9% improvement over the best baseline). In a post-hoc analysis, we also study the behavioral aspects of the targets and account holders and try to understand the reasons behind the incivility incidents. Interestingly, we observe that there are strong signals of repetitions in incivil behavior. In particular, we find that there are a significant fraction of account holders who act as repeat offenders - attacking the targets even more than 10 times. Similarly, there are also targets who get targeted multiple times. In general, the targets are found to have higher *reputation scores* than the account holders.

CCS Concepts: • **Human-centered computing** → **Social media**; **Social content sharing**; *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: Incivility, cyberbullying, opinion conflicts, Twitter

ACM Reference Format:

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2018. Opinion Conflicts: An Effective Route to Detect Incivility in Twitter. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 117 (January 2018), 27 pages. <https://doi.org/10.1145/3274386>

1 INTRODUCTION

Twitter is one of the most popular online micro-blogging and social networking platforms. However, of late, this social networking site has turned into a destination where people massively abuse and act in an incivil manner. This trend is affecting the users mentally and as a result they tend to leave

¹<https://phys.org/news/2016-11-twitter-tool-curb-online-abuse.html>

This research has been performed when SKM and AC were at IIT Kharagpur, India.

Authors' addresses: Suman Kalyan Maity, Northwestern University, Kellogg School of Management and Northwestern Institute on Complex Systems, Evanston, IL, 60208, USA, suman.maity@kellogg.northwestern.edu; Aishik Chakraborty, McGill University, School of Computer Science, Montreal, Canada, chakraborty.aishik@gmail.com; Pawan Goyal, Indian Institute of Technology Kharagpur, Department of Computer Science and Engineering, Kharagpur, 721302, India, pawang@cse.iitkgp.ernet.in; Animesh Mukherjee, Indian Institute of Technology Kharagpur, Department of Computer Science and Engineering, Kharagpur, 721302, India, animeshm@cse.iitkgp.ernet.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

2573-0142/2018/1-ART117 \$15.00

<https://doi.org/10.1145/3274386>

Twitter and other such social networking sites. In an article in Harvard Business Review 2016[44], the author claims that abuse and bullying are the primary reasons why this micro-blogging site is losing its active user base. For instance, in August 2014, the Gamergate controversy² [19, 20, 43, 68] broke out in Twitter and other social media platforms. The Gamergate controversy originated from alleged improprieties in video game journalism and quickly grew into a larger campaign (conducted primarily through the use of the hashtag #GamerGate) centered around sexism and social justice. Gamergate supporters took to massive incivil behavior e.g., sexual harassment, doxing, threats of rape and murder.

1.1 What is incivility?

In general, incivility involves sending harassing or threatening messages (via text message or e-mail), posting derogatory comments about someone on a website or a social networking site (such as Facebook, Twitter etc.), or physically threatening or intimidating someone in a variety of online settings [15, 28, 42, 65, 66, 78]. In contrast, cyberbullying is defined in the literature as intentional incivil behavior that is repeatedly carried out in an online context against a person who cannot easily defend himself or herself [63, 80, 105]. The important differences between incivility and cyberbullying are that for the latter there is i) a continuous repetition of the act and ii) the existence of imbalance of power between the target and the perpetrator.

1.2 Impact of incivility/bullying

Since online content spread fast and have a wider audience, the persistence and durability can make the target as well as bystanders read the account holder's words over and over again resulting in strongly adverse effects [16]. It can, thereby, potentially cause devastating psychological setbacks like depression, low self-esteem, suicide ideation, and may even ultimately lead to actual suicides among the targets [71] [50, 73]. Teenagers are mostly affected by this; in fact, more than half of the American teens have been the targets of incivility/bullying³. Not only kids, adults too are subjected to incivility/bullying⁴, ⁵. Owing to the increasing prevalence of social media, celebrity bullying also takes place frequently which is supported by various articles [9, 27, 30]. Facebook, Twitter, YouTube, Ask.fm, and Instagram have been listed as the top five networks having the highest percentage of users who report incidents of incivility⁶.

The rapid spread of this activity over the social media calls for immediate active research to understand how incivility occurs on OSNs today [40]. This investigation can help in developing effective techniques¹ to accurately detect and contain cases of incivility.

1.3 Working definition of incivility used in this paper

In this paper, we are specifically interested to study in depth the incivil behavior of Twitter users. Note that we do not consider cyberbullying in this study since it is very different in characteristics from incivility and calls for a completely separate line of investigation.

In the following we present a working definition of incivility that is largely accepted in the community and shall be used all through in the rest of the paper.

Working definition: We adopt the definition of incivility as an *act of sending or posting mean text messages* intended to mentally hurt, embarrass or humiliate another person using computers,

²https://en.wikipedia.org/wiki/Gamergate_controversy

³<https://cyberbullying.org/>

⁴<https://cyberbullying.org/bullying-is-not-just-a-kid-problem>

⁵<https://cyberbullying.org/Preventing-cyberbullying-top-ten-tips-for-adults.pdf>

⁶<https://www.ditchthelabel.org/research-papers/>

cell phones, and other electronic devices [28, 31, 32, 42, 94]. In the dataset section we precisely operationalize this definition for the experiments that follow.

1.4 Research objectives and contributions

In this paper, we analyze a large Twitter dataset for incivility detection followed by a detailed post-hoc analysis of the detected tweets. Toward this objective, we make the following contributions.

- We study the behavioral aspects of the targets and account holders and try to understand the reason for the incivility incidents. Our central observation is that *incivility in Twitter is strongly correlated to opinion conflicts between the account holder and the target*. Further analysis of target and account holder profiles across the linguistic and the cognitive dimensions reveals that account holders generally tend to use *more swear words, negations; express more emotions* and tweet more related to *body* and *sexual* categories.
- Once we have established the association of opinion differences between account holder and target toward named entities in incivility incidents, we propose a deep learning framework based on bi-directional LSTMs and character-CNNs and incorporate the entity-specific sentiment and followership information. Our model achieves an accuracy of **93.3%** with an F1-score of **0.82** which significantly outperforms (4.9%, 6.5% improvements w.r.t accuracy, F1-score respectively) the best performing baseline.
- We then conduct a post-hoc analysis of the incivility tweets on the entire dataset and study *repetitions in incivility*. We find that there exists *a significant fraction of account holders who act as repeat offenders*. In line of previous works [35, 36, 81], our results also confirm prior findings of targets being attacked multiple times and targets having usually more followers. The existence of this *imbalance of power* in terms of the social prestige and reputation between the target and the account holder is quite interesting because of its role reversal unlike cyberbullying incidents where the power lies mostly with the account holder.

1.5 Outline of the paper

We have various subsegments in this paper that we build up layer by layer to finally obtain the detection model. The entire pipeline is illustrated in Figure 1. As a first step tweets are labeled as incivil and then the timelines of the account holder and the target are crawled. This is followed by the construction of the *incivility context* (to be defined later). From the context, target sentiments are extracted using standard Target Dependent Sentiment Analysis (TDSA) technique. Opinions of the account holders and the targets are compared to finally ascertain opinion conflicts. This feature is then pipelined into to a deep neural framework to automatically classify a tweet as incivil.

2 RELATED WORK

There have been several works in the area of incivility - online harassment, cyberbullying, trolling etc., by researchers from various communities encompassing sociologists, psychologists and computer scientists. The overall literature in this area can be classified into two major categories based on the approaches taken to address the issues in incivility – i) survey/interview based approaches (non-computational) ii) computational approaches.

2.1 Survey/interview based approaches

Some of the very early works in the area of cyberbullying came from developmental psychology and sociology using survey/interview based approaches. Most of these studies deal with cyberbullying among school students and youth.

Prevalence of bullying: The prevalence of cyberbullying incidents in these studies varies from ~ 6% to 72%. Finkelhor et al. [39] report that ~ 6% of youth have been harassed online whereas

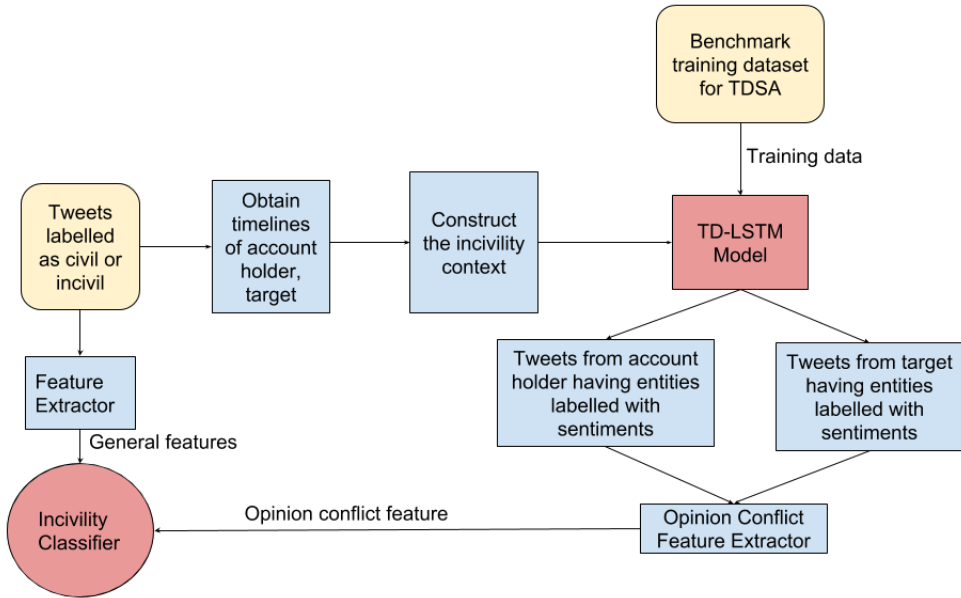


Fig. 1. Schematic of the steps for incivility detection. The yellow colored blocks represent inputs, the red colored blocks represent the classifiers and the blue colored blocks represent the intermediate steps.

Juvonen and Gross [59] report this number to be 72%. Most of the other studies estimate that 6 – 30% of teens have experienced cyberbullying [47]. Similarly, the number of youth who admit to cyberbullying is relatively lesser (3-44%) [47]. These variability in the numbers is largely due to the following facts – cyberbullying has been defined in various different ways [73, 99, 105] and also sometimes the perception of bullying differs across age (youth understand it differently compared to adults [11, 67]); the authors in these studies adopted various different sampling and methodological strategies [97]. Despite differences in rate of cyberbully incidents, the prominence of the factor among adolescents is well-grounded.

Sociodemographics: Cyberbullying though similar to traditional (offline) bullying embarks characteristic differences in sociodemographics [90]. In contrast to school bullying (offline) where boys are found to be more likely targets [17, 76], in cyberbullying there is no consensus on the role of gender in bullying [97]. Some studies have found that girls are more likely to be targets of cyberbullying [62, 100] yet other studies have found no gender differences [49, 102, 106]. Age is another characteristic in which cyberbullying patterns differ from traditional bullying. Although there is a decreasing trend of traditional bullying from middle to high school [76], some studies suggest that cyberbullying incidents increase during the middle school years [62, 102] and others have found no consistent relationship between cyberbullying and age [59, 95].

Effect of cyberbullying: There have been studies on the effect of cyberbullying incidents on the adolescents. Researchers have found that cyberbullying incidents are associated with experience of sadness, anger, frustration, depression, embarrassment or fear [48, 74, 79, 107] and these emotions have been correlated with delinquency and interpersonal violence among youth and young adults [2,

12, 69, 70]. Apart from these, cyberbullying has been associated with various other behavioral and psychological outcomes like suicide, school drop-out, aggression and fighting, drug usage and carrying weapons to schools [37, 48–51, 91, 106, 107]. In a recent study by Singh et al. [94], the authors have observed the influence of newer mobile app features like perceived ephemerality, location-based communication and image-based messaging etc. on cyberbullying in high school and its effect on school students.

Critical observations from the above: Interpretation of cyberbullying seems to be conditioned on age (youths and adults interpret them differently). The role of gender in cyberbullying is unclear. There could be a multitude of effects of cyberbullying ranging from depression, sadness, embarrassment to more severe incidents like fighting, drug use and carrying weapons.

2.2 Computational approaches

There have been several works in computer science domain mainly focusing on automatic detection of cyberbullying in social media largely based on text analytic approaches applied to online comments [22, 31, 61, 85, 89, 104, 108].

Content based approaches:

Language usage: There are several works that focus on usage of language - specific high frequent terms associated with bullying incidents. Reynolds et al. [85] use curse and insult words with their level of offensive intensity as primary indicators for detection of cyberbully incidents. Chen et al. [22] propose a user-level offensiveness detection method by introducing Lexical Syntactic Feature (LSF) architecture to detect offensive content and potential offensive users. They observe the contribution of pejoratives/profanities and obscenities in determining offensive content, and propose hand-crafted syntactic rules to detect name-calling harassments. They incorporate user's writing style, structure and specific cyberbullying content as features to predict the user's potential to post offensive content. Kontostathis et al. [61] also focus on language used in cyberbullying incidents. They perform analysis of the words used in connection to cyberbullying incidents on Formspring.me and further use those words and their context words to build a cyberbullying detection framework.

Gender role: Chisholm in a social study [25] show that there exist differences between males and females in the way they bully each other. Females tend to use relational styles of aggression, such as excluding someone from a group and ganging up against them, whereas males use more threatening expressions and profane words. Following this study, Dadvar et al. [29] leverage gender-specific language usage for cyberbully detection on MySpace profiles and show that such gender information improves the accuracy of the classifier.

Sentiment usage: Hee et al. [98] use sentiment lexicon and content of the text as bag-of-words (unigram/bigrams of words and character trigrams) for detection of cyberbully posts on a corpora of ~ 91,000 Dutch posts from Ask.fm. Nahar et al. [75] use sentiment features generated from Probabilistic Latent Semantic Analysis (PLSA) on cyberbully texts and bag-of-word features to detect cyberbullying incidents. Further, they detect and rank the most influential persons (bullies and targets).

Mitigation of cyberbullying incidents: Dinakar et al. [31] address the problem of detection of cyberbullying incidents and propose an intervention technique by notifying participants and network moderators and offering targeted educational material. In this work, they present an approach for bullying detection based on natural language processing and a common sense knowledge base that allows recognition over a broad spectrum of topics in everyday life. They construct BullySpace, a common sense knowledge base that encodes particular knowledge about bullying from various associated subject matters (e.g., appearance, intelligence, racial and ethnic slurs, social acceptance, and rejection) and then perform a joint reasoning with common sense knowledge. To mitigate the

problem of cyberbullying, they propose a set of intervention techniques. They propose an “air traffic control”-like dashboard, that alerts moderators to large-scale outbreaks of bullying incidents that appear to be escalating or spreading and help them prioritize the current deluge of user complaints. For potential victims, they provide educational material that informs them about coping with the situation, and connects them with emotional support from others.

Leveraging contextual information:

Apart from the content of the text, the contextual information is also important and relevant for cyberbully detection since the lexicon based filtering approach is prone to problems around word variations and lack of context. Yin et al. [108] use a supervised learning methodology for cyberbully detection using content and sentiment features, as well as contextual (documents in the vicinity) features of the considered documents on Slashdot and MySpace dataset. Similarly, Zhong et al. [110] study cyberbullying of images on Instagram using text as well as image contextual features. Hosseinmardi et al. [55] examine the users who are common to both Instagram and Ask.fm and analyze the negativity and positivity of word usage in posts by common users of these two social networks. Hosseinmardi et al. in two related works [56, 57] study detection and prediction of cyberbullying. In [57], the authors detect cyberbullying in Instagram by classifying images to different categories and further including text features from comments. In [56], the authors try to predict cyberbullying instances using the initial posting of the media object, any image features derived from the object, and any properties extant at the time of posting, such as graph features of the profile owner. In a recent paper [19], Chatzakou et al. employ a more robust approach considering text, user and users’ follower-followee network-based attributes for detecting aggression and bullying behavior on Twitter. In this paper, the prime objective of the authors is to distinguish cases of cyberbullying from aggression and spamming. In a related vein, Chatzakou et al. [18, 20] study cyberbullying and aggression behavior in GamerGate controversy (a coordinated campaign of harassment in the online world) on Twitter.

Kwak et al. [64] have analyzed a large-scale dataset of over 10 million player reports on 1.46 million toxic players from one of the most popular online game in the world, the League of Legends. They observe reporting behavior and find that players are not engaged in actively reporting toxic behavior and this engagement can be significantly improved via explicit pleas from other players to report. There are significantly varying perceptions of what constitutes toxic behavior between those that experienced it and neutral third parties. There are biases with respect to reporting allies vs. enemies. There are also significant cultural differences in perceptions concerning toxic behavior.

Chen et al. [21] in a recent study, present a dataset of user comments, using crowdsourcing for labeling. Due to ambiguity and subjectivity in abusive content from the perspective of individual reader, they propose an aggregated mechanism for assessing different opinions from different labelers. In addition, instead of the typical binary categories of abusive or not, they introduce an additional third category of ‘undecidedness’ to capture the instances that are neither blatantly abusive nor clearly harmless. They evaluate against the performance of various feature groups, e.g., syntactic, semantic and context-based features which yield better classification accuracy. Samghabadi et al. [88] perform similar study of nastiness (invective in online posts) detection in social media. They present evolving approaches for creating a linguistic resource to investigate nastiness in social media. The starting point is selecting profanity-laden posts as a likely hostile source for invective potentially leading to cyberbullying events. They use various types of classic and new features, and try to combine them for distinguishing extremely negative/nasty text from the rest of them. In a recent study, Hosseini et al. [53] analyze the recent advancement of Google’s Perspective API for detecting toxic comments and show that the system can be fooled by slight perturbation of abusive phrases to receive very low toxicity scores, while preserving the intended meaning.

Pavlopoulos et al. [81] have recently introduced deep neural models to moderate abusive (hate speech, cyberbullying etc.) user content.

Critical observations from the above: Both content and context have been extensively used to design hand-crafted features for detection of cyberbullying. There have also been one or two attempts to use deep learning techniques to abusive content moderation in general. However, there are hardly any approach that marry deep neural models with features that could be critically responsible for the invocation of incivil posts.

Hate Speech: Hate speech detection has been studied by various researchers. These works use several lexical properties such as n-gram features [77], character n-gram features [72], word and paragraph embeddings [33, 77] to detect hate speech. Apart from detection, there exist research works that look into various aspects of hate targets and the instigators. Silva et al. [93] study the targets of online hate speech by searching for sentence structures similar to “I <intensity> hate <targeted group>”. They find that the top targeted groups are primarily bullied for their ethnicity, behavior, physical characteristics, sexual orientation, class, or gender. ElSherief et al. [36] present the comparative study of hate speech instigators and target users on Twitter. They study the characteristics of hate instigators and targets in terms of their profile self-presentation, activities, and online visibility and observe that hate instigators target more popular (celebrities with a higher number of followers) Twitter users. Their personality analysis of hate instigators and targets show that both groups have eccentric personality facets that differ from the general Twitter population. In another concurrent paper, ElSherief et al. [35] focus on the hate targets - either directed toward an individual or toward a group of people. They perform the linguistic and psycholinguistic analysis of these two forms of hate speech and show that directed hate speech, being more personal and directed, is more informal, angrier, and often explicitly attacks the target (name calling) with fewer analytic words and more words suggesting authority and influence. Generalized hate speech, on the other hand, is dominated by religious hate, is characterized by the mentions of lethal words such as murder, exterminate, and kill; and quantity words such as million and many. In our study, as well, we observe similar findings regarding the accounts and targets.

Trolling: Trolling has been defined in the literature as behavior that falls outside acceptable bounds defined by those communities [7, 45]. There have been divided opinions on trolling behavior. Prior research works suggest that trolls are born and not made: those engaging in trolling behavior have unique personality traits [14] and motivations [3, 46, 92]. However, there is other school of thought suggesting that people can be influenced by their environment to act aggressively [26, 58]. Cheng et al. [23] in a recent study, focus on the causes of trolling behavior in discussion communities. By understanding the reason behind trolling and its spreads in communities, one can design more robust social systems that can guard against such undesirable behavior.

2.3 The present work

Our work is different from the previous works in several ways. We focus on understanding the incivility incidents on general population in Twitter (unlike most of the previous studies which are based on children or teens and dated online platforms like FormSpring, MySpace etc.). While most of the survey/interview based techniques concentrate on analyzing the effects and consequences of these incidents after the incident has happened, we attempt to (i) early detect the incidents of incivility in the first place and then (ii) analyze the behavioral aspects of the account holder and the targets so detected. Though there have been several lexicon-based studies, they have rarely reused any labeled data from previous researchers due to incompatibility issues of applying on different platforms. Our study, though on Twitter, can be used in other social media platforms because of presence of a single user in multiple platforms and essentially similar language usage trend.

In particular, we observe that opinion conflict (target and account holder showing opposite sentiments towards same named entities) is associated with incivility and we believe that this aspect has not been earlier reported in the literature, not even in any of the earlier computational studies. We then propose a deep learning based model that can automatically detect incivility text more efficiently by correctly exploiting the aforementioned connection between incivility incidents and target dependent expression of sentiments. This fusion of a deep neural model with a critical feature that stems from opinion conflict is a prime novelty of our work.

3 DATASET PREPARATION

Recall that our working definition of incivility refers to the act of sending or posting mean text messages intended to mentally hurt or embarrass a target. In this section we describe how we operationalize this definition through the assemblage of appropriate data.

3.1 Operationalizing the definition of incivility

Although incivility is rising in Twitter¹, from automated crawls, it is difficult to obtain data containing direct incivility instances. However, our working definition presents us with one important clue that incivility tweets should generally contain *offensive/mean* words. We use a list of such offensive words compiled by Luis von Ahn of Carnegie Mellon University⁷. This list is very comprehensive and larger than any list containing offensive words known to us or used previously by other authors. This list has been used by various authors [1, 4, 13, 21, 41, 83] in the past and this rich literature justifies our choice. The list contains 1374 offensive words. The offensive words included in the list can be categorized as: (i) Swear/profane words: *f**k*, *a\$\$hole*, *b*tch* etc., (ii) Negative words: *die*, *hell*, *death* etc. and (iii) Others: *enemy*, *drug* etc.

Note that the list contains certain words that some of the people would not find offensive. However, incivil tweets have a very high chance of having these offensive words⁸. We do understand that these may not cover all instances of incivility and might affect the recall of our models which is a limitation of our treatment of the problem. However, the precision of the system should not be affected by this limitation and the current target of the paper is to build a highly precise system.

We then use these words to filter the Twitter data stream. The crawling has been done from August to December 2017. We filter the tweets in which any of those offensive words are present. We obtain a total of ~2,000,000 tweets containing one or more such offensive words. Out of the 1374 offensive words, 59.4% of words have been used at least once, 54.9% have been used at least twice and 47.8% have been used at least thrice. Some offensive words have also been used in as high as 1.3% of the tweets.

3.2 Mention based filtering of the dataset

We further filter the dataset based on the presence of mentions in the tweets. This is because in general, any conversation in Twitter should contain mention(s) and an account holder would generally mention the target in their (incivil) tweets. We consider only those tweets in which one or more mentions appear. This reduces the tweet dataset to ~300,000 tweets from the earlier ~2,000,000 tweets. Note that this step operationalizes the second part of the working definition of incivility, i.e., the offensive text is usually targeted to an individual, which in this case is the mentioned target.

⁷<https://www.cs.cmu.edu/~biglou/resources/bad-words.txt>

⁸Note that the number of incivil tweets in a random sample of tweets is very low. In fact, we observe that in a random sample of 100 tweets, there are only 2 tweets that fall into the category of incivility. Also the tweets that were incivil from the random sample contained words from the “mean” words list. After obtaining manually annotating the tweets as incivil or civil, we found that the fraction of mean words were higher in the incivil tweets.

3.3 Manual labeling of the tweets

While the previous step reduces some noise in the data, not all the tweets obtained are related to incivility. We, therefore, randomly sample 25,000 tweets out of these ~300,000 tweets and manually label them as incivil/civil instances. We consider those cases where we have full agreement between the two authors who did the labeling. Out of the 25,000 tweets, we discard 729 tweets where there was a disagreement between the authors. In the 24,271 selected tweets where there was full agreement, we find 8,800 tweets as instances of incivility and the remaining 15,441 tweets as instances of civility.

3.4 Criteria for labeling the tweets

Following are some of the broad cases where we have labelled the tweets as being incivil:

- (a) **Blackmails or threats:** These tweets have expressions of physical or psychological threats to the targets. Example: ‘I’ll smash you in the face when I see you.’
- (b) **Insult:** These tweets have insults that are abusive for the target. Example: ‘You are such an a\$\$hole.’
- (c) **Cursing:** These tweets have expressions wishing that some grave misfortune befalls the target (like their or their loved one’s death). Example: ‘You’ll die and burn in hell.’
- (d) **Sexual harassment:** These tweets contain unwanted sexual talk which might be derogatory. Example: ‘Post a naked pic u sl*t!’

While labelling the tweets, we took into account that some of these tweets although might contain offensive words are not cases of incivility, eg., ‘Hey bitches, feel like seeing a movie tonight?’

4 TARGET SENTIMENTS & INCIVILITY

In this section, we construct an *incivility context* to be able to perform an in-depth analysis of the reason for incivility. We find that contradicting sentiments (opinion conflicts) by the account holder and target toward a named entity, are highly correlated to the act of incivility.

4.1 The incivility context

As described in the previous section, we label the incivil tweets and from the mention relationship, identify the target and the account holder. A typical example is as follows:

@user1 what a dumb ass ** person you are, user2
Here user2 is the account holder while user1 is a target.

In our work, we consider the mentions of an incivil tweet as the target. In case of multiple mentions, we consider all of the mentions as targets. However, it might be possible that some of the mentions are not actually targets (usually rare as observed through manual inspection) and is currently a limitation of our work which we wish to address in the future. Once we have the account holder-target pairs, we consider the targets’ and the account holders’ profiles and crawl their timeline (we get a maximum of 3,200 tweets per user). Also note that the timeline is crawled at around the same time we collected our dataset. We take the first 100 tweets from their timeline. We find that taking the first 100 tweets itself contains ample clues for us to investigate. The intuition behind constructing this context is that these tweets might contain the context that triggered the incivil tweet and thus these tweets should provide the best clues as to why the account holder targeted the target. In fact, for annotating the tweets in the previous section, the annotators in many instances had to visit the target and the account holder profiles to ensure if the instance was indeed a case of incivility. In almost all the cases where the annotators attested instances of incivility they observed opinion conflicts between the target and the account holder. We therefore attempt to automate this natural “back-and-forth” process adopted by the annotators for identifying

incivility instances and build the incivility context to automatically analyze such opinion conflicts by making comparisons between the target's and the account holder's timelines.

From the example of incivility context cited below, we observe that the target tweets positively about *Donald Trump* and *US Economy*. However, the account holder tweets negatively about *Trump* and positively about *President Obama*. We can observe that there is a conflict of opinion between the target and the account holder as the sentiments expressed toward the common named entity Donald Trump is opposite. Going through the entire context, we find that this opinion conflict leads to an incident of incivil post.

account holder's tweet:

@user1 Enjoy prison a\$\$hole!

account holder's context tweets:

@user5 @user6 You sir, are just another clueless Trump lemming.

@user7 @user8 Seriously, get your head out of Trump's ass already. Go watch your Fox News & Friends and eat your jello.

@user8 The video of what your boyfriend said: Trump labels US justice system 'laughingstock' @CNNPolitics
<https://t.co/QNa2jqAYsE>

@user9 if the Devil was running as a Republican, would you still vote for him? Your morals and priorities are so screwed up.

@user5 Seriously, let it f**king go. You are worse than a scorned girlfriend bringing up decades of shit that does not matter. You are the BIGGEST LOSER of all time.

@user11 Trump idiot lemmings are condemning the outrage over slavery and agreeing w/the idiot Kelly about praising Lee? Clueless losers

target's context tweets:

@user10 You are truly stupid. Trump is the first President to come into Office supporting marriage equality

Strange that the #fakenews media never gets stories wrong in favor of Trump. It's almost like they do it on purpose

According to HuffPo, President Trump is effective, but they don't like it. Donald Trump's relentless focus on tax cuts, deregulation and draining the swamp is great for job growth... with minorities

and so on ...

Based on the 8800 incivility incidents in our manually labeled dataset, we obtain the incivility contexts. Since our hypothesis is that opinion conflicts between the account holders and the targets create the rift between them, we plan to use sentiment analysis as a tool to automatically identify the conflicts. In particular, we try to find out the sentiments expressed by the targets toward different named entities in their tweets. We also find out the sentiments expressed by the account holders toward the different named entities. Next we find the common named entities toward which the target and the account holder expresses positive or negative sentiments. Finally, we identify if the targets and the account holder have expressed opposing sentiments toward the same entity. We propose an algorithm to detect such opinion conflicts easily in the following subsections.

4.2 Step 1: Named entity recognition

We use a named entity recognition (NER) system to identify named entities that will enable us to proceed further with the sentiment analysis. Recall that, according to our hypothesis, we first need to identify the named entities and then find the sentiments expressed by the users toward these named entities. For this purpose, we use the tool proposed in [86, 87]⁹ that is trained on Twitter data and performs much better than other traditional NER tools like the ones proposed in [38]¹⁰.

⁹https://github.com/aritter/twitter_nlp

¹⁰<http://nlp.stanford.edu/software/CRF-NER.shtml>

Classifier	Accuracy	Macro F1
TD-LSTM on benchmark dataset	69.2%	0.68
TD-LSTM on our dataset	73.4%	0.71

Table 1. Accuracy of the TD-LSTM.

4.3 Step 2: Target dependent sentiment analysis

After named entity recognition, we set out to obtain the sentiment polarity associated with each named entity in a given text. Our sentiment analysis model is based on an LSTM framework. A brief description of LSTMs and how we use it for target sentiment detection follows below.

Long Short Term Memory (LSTM) Recurrent Neural Networks (RNNs) are used for handling tasks that use sequences. A recurrent network takes an input vector x and outputs a vector y . However, y depends on not only the current input, but on all the inputs fed into it in the past. RNNs are however not capable of handling long term dependences in practice [5]. It was observed that the backpropagation dynamics caused gradients in an RNN to either vanish or explode. The exploding gradient problem can be solved by the use of gradient clipping. [52] introduced LSTM to mitigate the problem of vanishing gradient. The LSTMs by design have a hidden state vector (h_t) and also a memory vector c_t at each timestep t . The gate equations at timestep t with input as x_t and output as o_t are:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$C_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ C_t$$

$$h_t = o_t \circ \tanh(c_t)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Here, the three vectors f, i, o are thought of as binary gates to control whether each memory cell is updated. The \circ operation here denotes the element-wise matrix multiplication.

Target dependent sentiment classification using LSTM: Here we describe a LSTM based model for target dependent (TD) sentiment classification. The model we use is adapted from [96]. The sentence representation in this model can be naturally considered as the feature to predict the sentiment polarity of the sentence. We use the glove vector word embeddings which are trained on 27 billion tokens and 2 billion tweets as the vector representation of the words. We use the target dependent LSTM approach (TD-LSTM) to find the sentiment polarity of the tweets toward a target.

The training of the TD-LSTM is done in a supervised learning framework using cross entropy loss. We perform all parameter updates through backpropagation and *stochastic gradient descent*.

Evaluation results: We use the benchmark dataset from [34]. We train it on their Twitter dataset having 6248 sentences and test it on 692 sentences. The percentages of positive, negative and neutral instances in both the training and test sets are 25%, 25% and 50% respectively. The accuracies reported in table 1 are obtained using 200 dimensional glove word embeddings. We manually label the sentiments of 100 random entities from our dataset. We achieve an accuracy of 73.4% on our dataset.

The above accuracy has been obtained using the early stopping criteria and the same weight initializations as mentioned in [96].

4.4 Observations from TD sentiment analysis

In this section, we analyze the outcome of the TD sentiment analysis on the named entities. For each incivility context, we perform the NER and target dependent sentiment analysis. Specifically, we take all tweets in the incivility context of the account holders and targets and then run the NER on the tweets to extract the named entities. Then we run the TD-LSTM with targets as these named entities to get the target’s polarities toward the named entities. We can conclude that the cases of incivility where the algorithm returns empty sets for both the positive and the negative sentiments for all the named entities are not related to contradicting sentiments between the account holder and the target.

Sentiment expression toward different named entities: In Figure 2, we show the distribution of

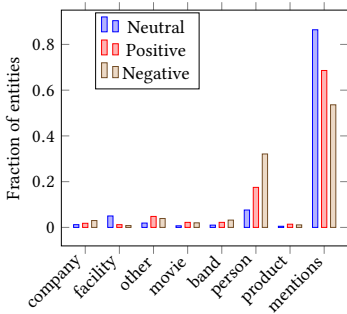


Fig. 2. Comparison of sentiment toward different named entity classes.

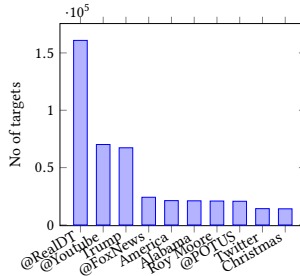


Fig. 3. Popular entities toward which different targets express sentiments. (@RealDonaldTrump has been shortened as @RealDT.)

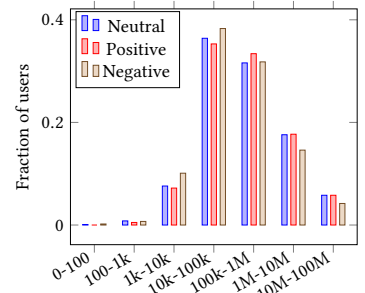


Fig. 4. Sentiments versus followership counts of targets.

sentiments expressed toward different named entity classes by the targets in the incivility context. It is evident from the figure that users express opinions differently about different entities. For example, many users express negative sentiments toward a mention or a person. However, users express no opinion toward the majority of mentions. Further, users express negative sentiments more often than positive sentiments. This observation can act as an effective strategy to predict incivility incidents beforehand.

Next we attempt to identify the popular entities toward which targets express sentiments (see Figure 3). It is evident from the figure that there are more targets expressing opinion about Trump, Youtube and Fox News. These entities might come on top due to the particular dataset used for this work.

Table 2 shows examples where the target is attacked for expressing positive sentiment toward Trump in an example and negative sentiment in another.

Sentiments and followership: We next study whether followership counts and the sentiments expressed by a target have any correlation. Figure 4 shows the relationship between the sentiment expressed by the users and their number of followers. There is a distinct trend observed here. Users with moderately low followership (100-10K) tend to express more negative sentiments toward the named entities while the users with high followership (100K-10M) show more positive sentiments toward such entities. Usually incivility is associated to negative sentiments. It seems that users with high followership are less likely to be an account holder. In the later sections, we show that this is indeed true.

Opinion conflicts. We have defined opinion conflicts as the case where the target and the account holder express opposing sentiments toward the same entity. For example, suppose the target T

Table 2. Sentiment toward a common named entity

Positive Sentiment	Negative Sentiment
Incivil tweets	
@user1 you really are a stupid dumb **. Do you know anything. Like trump you are a ** brainless **. Dumbo	@user2 @user3 @user4 Since when do you have to be a politician to talk to people with respect you stupid moron. You are a fool.
Incivility context	
RT @user5: The Washington Post calls out #Crooked-Hillary for what she REALLY is. A PATHOLOGICAL LIAR! Watch that nose grow! RT @user6: The Obama/Clinton admin has failed Americans: "The Obama economy is trouble for Hillary Clinton" https://t.co/JBoCmyHj6v	I don't think he's got a thing to apologize for.. He is not a professional politician @user7: Trump is trying to go negative; drive up those numbers so that he & Clinton are on the same page https://t.co/RIB00L32v9

expresses 10 sentiments toward entity E, out of which 8 are positive and 2 are negative. Similarly, suppose the account holder expresses 20 sentiments toward entity E, out of which 4 are positive and 16 are negative. Then, the overall sentiment toward the entity E by T is positive, whereas by A is negative. This counts as a single opinion conflict between A and T. We use this mechanism to compute the opinion conflict feature values.

We find that opinion sentiments are highly correlated to incivility. We observe that 75% of the incivil tweets in our dataset have at least *one* opinion conflict between the targets and account holders. In contrast, there are only 53% civil incidents where there is an opinion conflict. Further, for each incivil incident, there are 2.52 ± 0.03 similar sentiments expressed, whereas, for each civil incident, the mean similar sentiments expressed is 4.49 ± 0.19 . Thus, it is clear that most civility contexts come with an agreement of opinion (67% of total sentiments) whereas an incivility context is filled with opinion conflicts (49% of total sentiments). Therefore, we use this concept as a potential strategy for incivility detection and further use this as a feature in our subsequent models.

5 BEHAVIOR OF THE ACCOUNT HOLDERS AND THE TARGETS

In this section, we study the socio-linguistic behavior of the targets and the account holders. Specifically, we analyze their user profiles to study followership behavior, tweeting behavior and psycholinguistic dimensions.

5.1 Followerships

In Figure 5, we show the followership distribution of the user profiles in our dataset. It is evident from the figure that our dataset is a good mix of both normal users as well as celebrity users. Figure 6 shows a comparison between the follower counts of the account holder and the target accounts. The figure indicates that the target accounts are more popular than the account holder accounts (p -value of significance is $< 10^{-5}$). Apparently, many well known users or celebrities tend to be targets of incivility. A number of news articles also seem to mention this [9, 27, 30]. Moreover, usually celebrities (i.e., those with typically high follower counts) tend to avoid insulting someone on social media sites.

5.2 Tweeting behavior

The following analysis has been done on the timeline crawls of the account holders and the targets to identify differences in their respective tweeting patterns. Figure 7 shows a comparison between the account holder and the target accounts with respect to their tweeting behavior. We observe that there are more account holders who tweet less or moderately compared to the targets (p -value of significance is $< 10^{-5}$). However, in the high tweeting zone, there are comparatively more targets than the account holders. In fact, there are several accounts with more than 100K tweets. These are

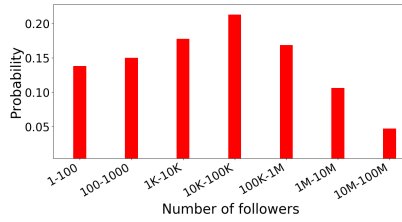


Fig. 5. Followership distribution of all the users in our dataset.

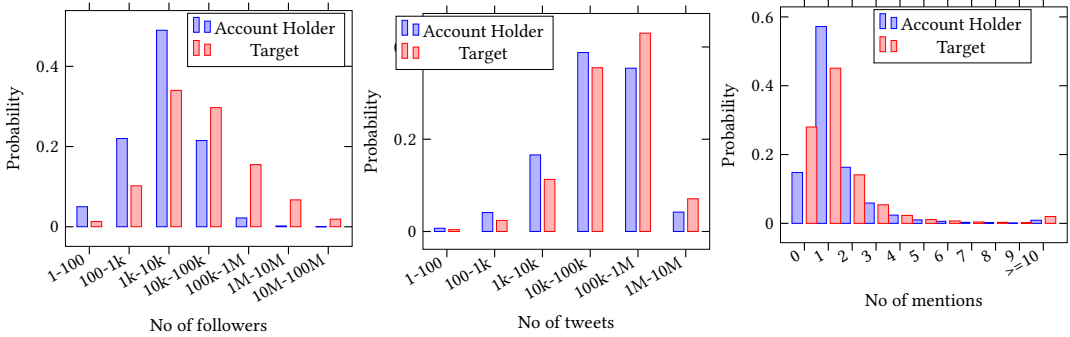


Fig. 6. Comparison of number of followers of the account holders and the target accounts.

Fig. 7. Comparison of the number of tweets posted by the account holder and the target accounts.

Fig. 8. Comparison between number of mentions in a tweet by the account holder vs the target accounts.

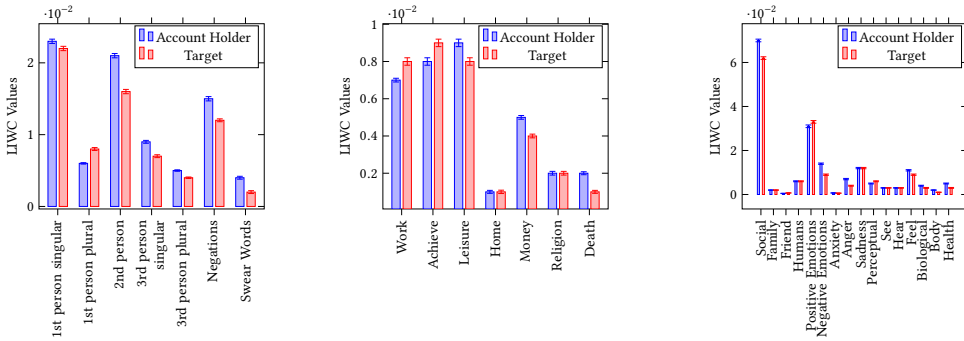
typically news media accounts like Fox News, Sky Sports, Telegraph, CBS Sweden, CNN News 18 etc. They are being cyberaggressed for the news/opinions the news anchors associated with them typically express. We cite below such an example of a news media involved in incivility.

@user1 @newsmedia1 stupid English b*tch asking is there people on that plane?... No you thic k c*nt it's like google cars...
Incivility context:
 A Dubai firefighter has died of injuries sustained putting out fire after plane crash landing - Emirates chairman
<https://t.co/i25sjAKfOC>

We also observe the mention patterns of the targets and account holders. Figure 8 shows the distribution of number of mentions in a tweet by the account holders versus the distribution of number of mentions in a tweet by the targets. It is evident that there are many targets (~ 30%) who do not mention at all. The account holders generally mention multiple people in a tweet and much more often than the targets (p -value of significance is $< 10^{-5}$). This also indicates that they might be insulting multiple targets. We further observe that there are some targets who use more mentions in their tweets than account holders. These targets have high followers (198K on average) and they probably use more mentions mostly for promotional purposes.

5.3 Linguistic and cognitive dimensions

In this section, we perform linguistic and psychological measurements of the tweets posted by the target and the account holder profiles. We use the Linguistic Inquiry and Word Count (LIWC) [82], a text analysis software to find which categories of words are used by the targets and the account holders. LIWC evaluates different aspects of word usage in different meaningful categories, by counting the number of words across the text for each category. Table 3 notes the frequent words



(a) Comparison of LIWC scores in the linguistic categories between the account holder and the target tweets.

(b) Comparison of LIWC scores in personal concerns categories between account holder and target tweets.

(c) Comparison of LIWC scores in the cognitive categories between the account holder and the target tweets.

Fig. 9. LIWC analysis. We have used standard error as error bars in the graph.

from our dataset corresponding to the different LIWC categories. In Figure 9a, we show the linguistic categories for the target and the account holder tweets. It is quite evident from the figure that the account holders use first person quite frequently compared to the targets and refer to others (here, targets) in third person. Moreover, the use of swear words and negations are more in the account holder tweets than the target tweets.

Next, we look at some of the personal concerns such as work, achievements, leisure, etc. Figure 9b shows the differences of the account holder and the target profiles in terms of expression of personal concerns. We observe that the targets tweet more about work and achievements; account holders tweet more about monetary aspects. We also observe that in “religion” (e.g., church, mosque) and “death” (e.g., bury, kill) categories, account holders tweet more than the targets. This is indicative of the fact that religion-based incivility is more prevalent in social media. This is in line with the work by Hosseinmardi et al. [54] where they found that there is a high usage of profane words around words like “muslim” in ask.fm social media.

We further analyze the cognitive aspects of the target and the account holder tweets (see Figure 9c). We observe that account holders are more expressive of their emotions than the targets. Further, the account holders tend to tweet more related to “body” (e.g. face, wear) and “sexual” (e.g. sl*t, rapist etc.) categories. In “friend” category, both account holder and target tend to tweet in similar proportions. However, the account holders talk more frequently related to the “social” category. Also targets express more “positive emotion” whereas, account holders express “negative emotion”.

6 DETECTING INCIVILITY

In the earlier section, we have discussed the possible connections of incivility with target dependent sentiments and observed the various behavioral aspects of the targets and the account holders. Note that to do this, we had to manually label the incivil tweets as a first step in order to build the incivility context. In this section, we shall propose an automatic approach to classify incivil tweets. **Training and test sets:** Recall that we have 24,271 manually labeled tweets. We choose randomly 21,000 of these tweets and consider them as our training set. The remaining 3271 points are in the test set.

Table 3. Frequent words from the LIWC category in our dataset. The asterisk here denotes the acceptance of all letters, hyphens, or numbers following its appearance.¹¹

Category	Frequent words from the dataset
Swear words	shit*, dumb*, bloody, crap, fuck
Work	Read, police, political, policy, student*
Achieve	Better, win, won, first, best
Leisure	party*, read, running, show, shows
Home	clean*, address, home, family, house*
Money	Free, money*, worth, trade*, tax
Religion	Sacred, moral, worship*, hell, devil*
Death	War, death*, murder*, kill*, die
Social	You, we, your, our, they
Family	Family, families*, pa, mother, ma
Friends	Mate, mates, fellow*, lover*, friend*
Humans	people*, human*, women*, children*, woman
Positive Emotions	Like, party*, lol, better, support
Negative Emotions	War, wrong*, violent*, liar*, rape*
Anxiety	doubt*, fear, risk*, avoid*, afraid
Anger	War, violent*, liar*, rape*, fight*
Sadness	low*, lost, lose, loser*, fail*
Perceptual	green*, say*, said, watch*, see
See	green*, watch*, see, white*, look
Hear	say*, said, hear, heard, listen
Feel	round*, hard, loose*, hand, feel
Biological	health*, drug*, rape*, life, shit*
Body	shit*, head, brain*, hand, face
Health	health*, drug*, life, weak*, living

6.1 Baseline models

There have been several studies like [10] that consider identifying incivility instances in various social media platforms based on various content features. We adopt this work (baseline 1) for comparison with our model. A few studies use basic n-gram features. Hosseinmardi et al. [57] uses unigrams and tri-grams to detect incivility. Xu et al. [104] uses unigrams, unigrams+bigrams as features to detect incivil traces in social media. These n-gram features constitute our second baseline model. We also consider n-gram (uni-, bi-,tri-) based model with automatic feature selection to reduce the feature space to smaller number of features. We also use the work by Chen et al. [21] as a baseline. In particular, we use n-grams together with textual features. The textual features include number of words in the tweet, number of characters in the tweet, number of sentences in the comment, average word length (#characters divided by #words), average sentence length (#words divided by #sentences), profane words usage level (#profane words divided by #words), uppercase letter usage (#uppercase letters divided by #sentences), punctuations usage (#punctuations divided by #sentences), URL usage level (#URL divided by #words), mentions usage level (#mentions divided by #words) For all the above models, we use different classifiers like SVM with linear and rbf kernels, k-NN, logistic regression, Adaboost with L2 regularization wherever possible and report the best overall accuracy.

Feature engineering: We use the following content features adopted from Bommerson [10].

Length of the tweet. This feature simply corresponds to the length of the tweet measured in terms of the number of words in the tweet.

Number of negative/offensive words in the tweet. This feature calculates the number of offensive words using the offensive words list we introduced earlier.

Severity value of the tweet. We label every word in the offensive words list manually as ‘1’ or ‘2’ and this is unanimously agreed upon by two of the authors. This is done because all offensive words are not equally bad; some have more severe effect when used in the tweet. We label the more severe offensive words as ‘2’ while the less severe ones are labeled as ‘1’. If w is a bad word in a tweet and W is a set of all offensive words, then severity of a tweet = $\sum_{w \in W} sev(w)$, where $sev(w)$ denotes the severity of the word w and $sev(w) \in \{1, 2\}$.

Time of the tweet. The time of post of a tweet is taken as a feature. It can be a distinguishing feature between the two classes as incivility among teenagers happens usually after school whereas normal tweets flow in all throughout the day.

Negation used or not. Sometimes the offensive words can be used with negations. This however reverses the effect of the word and thus the tweets in these cases are usually civil (e.g., '@xyz Please don't die.'). To handle this case, we take into account if negations are used along with offensive words.

Drawbacks of the baseline models: The baseline model suffers from the following drawbacks:

Sentence dependencies. The above features cannot detect dependencies between different sentences in a tweet. For example, '@xyz you are a footballer. All footballers are brainless.'

Multiple connotations. The above features do not take into account the multiple connotations of a word or phrase. For example, the word 'well' has a neutral connotation whereas the words 'sooo welllll' indicates a much happier/excited state of the mind of the author.

Obfuscated words. The account holder can use words like 'a\$\$hole' or 'f**k' which will not get detected by the above features.

In this work, we use character-level models that build representations of sentences using the constituent characters. The advantage of using these kinds of models is that we can easily deal with non-traditional spellings used in social media website and hence, can easily capture obfuscated words, multiple connotations of words and sentence dependencies without the need of normalizing the words to their correct forms. In particular, we use several deep learning models that take such character embeddings as input and output a sentence representation using which we are able to classify if the tweet is related to incivility or not. As an additional deep learning based baseline we use the model proposed by Pavlopoulos et al. [81] that was introduced to moderate abusive user content.

6.2 The deep learning frameworks

We propose deep learning frameworks primarily based on character-level LSTM and character-level CNNs. We use character-level models in this experiment as the characters can provide more information than the words as a whole for the tweet dataset under consideration. In particular, since the tweets are limited to only 140 (now 280) characters, it is very difficult to obtain rich embeddings from such a short text where only a few words are present. Therefore, character-level LSTM/CNN can obtain more information than a word-level LSTM/CNN in such a scenario. We would also like the model to have knowledge about the past sequence of characters as well as to look into the future. Thus, we choose the bidirectional LSTM framework¹² for our purpose.

The second model we use is a character-level CNN model. The authors in [109] have successfully used character level CNNs for text classification. This motivated us to try CNNs for the purpose of our task. We use a model with 2 convolution layers, each layer followed by a max-pooling layer.

Bi-directional LSTM: We use a character-level bi-directional LSTM with *tanh* as the function that provides non-linearity. The 100 dimensional characters embeddings are initialized randomly and updated during training. The encoded vectors are passed to the LSTM units and then the output of each unit which is a 100 dimensional vector is passed to the next unit in the forward as well as backward direction. We concatenate the output of the LSTM unit in the last timestep in both forward and backward direction. Finally, we use a softmax layer to calculate the probability of the tweet belonging to a particular class.

Character CNN: The second model we use is a character-level CNN model with ReLU non-linearity. The first convolution layer contains 100 filters with kernel size 5×5 . The second convolution layer,

¹²We have also compared the model with unidirectional LSTM, GRUs although we do not report the results for all these.

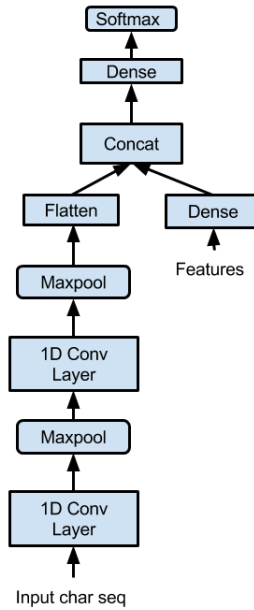


Fig. 10. char-CNN + opinion conflict feature based model for incivility detection.

once again, contains 100 filters with kernel size 5×5 . Both the convolution layers are followed by a max pooling layer. Finally, we flatten the output vector and use a dense layer followed by a softmax to get the output probabilities for the two classes. We also use dropout and batch normalization in both the above models to prevent overfitting.

Character CNN + opinion conflict feature: The char-LSTM and char-CNN models might be able to extract some features and temporal relations between characters but still they might not be able to take into account the crucial observations about incivility that we made in the two previous sections. To improve the performance of the model and to take advantage of both the scenarios, in the char-CNN model, we fuse the flattened vector we get after the two convolution and maxpool layers with a special entity sentiment based feature, which we have seen in the previous sections to be very discriminative. The full architecture is shown in Figure 10. In particular, we take the vector obtained by flattening and concatenate it with the *opinion conflict feature* described below. Then we take this vector as an input to a feed forward neural network which classifies the tweet as an instance of incivility or otherwise. Thus, we train the fused model end-to-end according to the final loss function. The weights learned from the feed forward network at the end of the concatenated flattened vector and custom feature vector tell the model the amount by which each feature should be weighed.

Opinion conflict feature: The opinion conflict feature is the total count of opinion conflicts that appear in the incivility context between the account holder and all the targets in a particular tweet.

Training: We perform all parameter updates through backpropagation and *stochastic gradient descent*. We apply dropout with probability 0.25. We also use early stopping with patience value as 3.

Results: We train the model in a supervised learning framework using the same training and test data on which we had trained the baseline model. The results are noted in Table 4. The first observation is that content based features do not perform very well in this setting (see pink row).

Table 4. Classification results

Method	Accuracy	F1-Score	ROC-AUC
Content based features			
Bommerson et al. [10])	71.1%	0.27	0.61
n-gram based features			
unigrams	86.4%	0.73	0.91
unigrams (automatic feature selection)	85.6%	0.72	0.91
bigrams	86.7%	0.66	0.79
bigrams (automatic feature selection)	88.4%	0.69	0.81
trigrams	82.3%	0.18	0.54
trigrams (automatic feature selection)	83.6%	0.29	0.59
unigrams + bigrams	86.7%	0.74	0.91
unigrams + bigrams (automatic feature selection) (Xu et al. [104])	87.5%	0.67	0.79
unigrams + trigrams	86.7%	0.63	0.62
unigrams + trigrams (automatic feature selection) (Hosseinmardi et al. [57])	86.7%	0.64	0.62
bigrams + trigrams	86.8%	0.66	0.79
bigrams + trigrams (automatic feature selection)	88.8%	0.76	0.93
unigrams + bigrams + trigrams	86.8%	0.73	0.91
unigrams + bigrams + trigrams (automatic feature selection)	88.9%	0.77	0.92
baseline 1 + opinion conflict	76.1%	0.37	0.61
unigrams + bigrams + trigrams + opinion conflict	80.9%	0.77	0.90
unigrams + bigrams + trigrams + textual features (Chen et al. [21])	86.7%	0.77	0.92
char-LSTM and char-CNN Models ¹			
char-LSTM	88.9%	0.80	0.84
char-LSTM+attention (Pavlopoulos et al. [81])	78.5%	0.53	0.74
char-LSTM+attention (Pavlopoulos et al. [81])+opinion conflict	78.6%	0.54	0.75
char-CNN	93.0%	0.81	0.88
char-CNN + opinion conflict	93.3%	0.82	0.89

¹ Results have been obtained by taking mean of 10 random runs.

As we can see from the table, the character CNN model along with the opinion conflict feature (see green row) outperforms the baseline models achieving 93.3% accuracy with F1-score of 0.82 and ROC area of 0.89. Note that while the increase in F1-score from char-CNN to char-CNN + opinion conflict model is from 0.81 to 0.82, it is statistically significant with p -value < 0.05 . Furthermore the char-CNN + opinion conflict model corrects 10.14% of the errors made by the char-CNN model.

Among the baseline models, the best performance is obtained when we use unigram, bigram and trigrams as features (see yellow row). Our model significantly outperforms this best performing baseline (5.1%, 6.5% improvements w.r.t accuracy and F1-score). The CNN model seems to learn better representations than hand picked bag of word n-gram features.

Below we list a couple of examples where char-CNN fails and char-CNN+opinion conflict predicts correctly. Both cases have very few linguistic cues to help in the classification task. In the first case, the account holder tweets negatively about Old Alabama whereas the target @user1 tweets positively about Old Alabama. In the second case, we found that the account holder tweets positively about Daily News and @user1, the target negatively about Daily News.

1. @user1 @user2 YOU and your incessant attempts at lying to and bullying the press while at the White House as PS are reason for apology you sack of dog stools. The AMERICAN public does not distrust the media other than by your disingenuous propaganda. Ur time is coming you asshole.

2. @user1 No just you. Everyone would know he was asking for you. Asshole!

7 POST-HOC ANALYSIS

We run the trained char-CNN model on the entire dataset of $\sim 300,000$ tweets and perform a post-hoc analysis of the incivility tweets to extract some more interesting properties. The key observation is that there are multiple instances of incivil posts from the same account holder as well as a target being attacked multiple times.

7.1 Repetition of incivility and targeting

In this section, we analyze the incivil posts in more detail focusing on their repetitions. Following are some examples of multiple incivility and targeting. The first two examples show @user1 being attacked by multiple account holders (user3 and user4) and the last two examples show @user5 being attacked multiple times by user6.

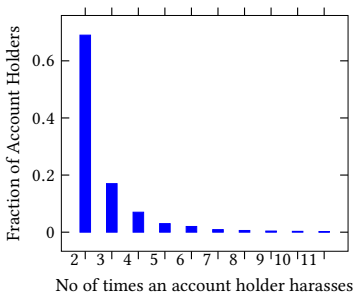


Fig. 11. Account holders involved in multiple instances of incivility.

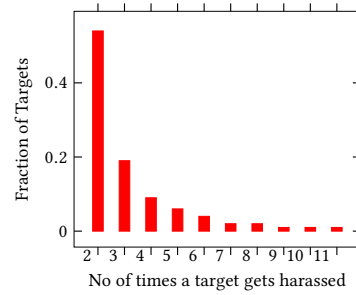


Fig. 12. Targets attacked multiple times.

@user1 Hillary was a skankb*tch,and they make a cream for your butt-hurt condition (account holder: user3)
 @user1 Get ur a\$\$ off here, u stupid b*tch (account holder: user4)

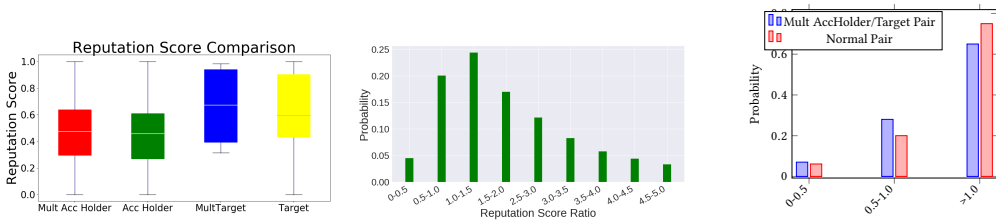
@user5 you are the biggest f**king piece of scum there is. Karma is a b*tch (account holder: user6)
 @user5 you are the biggest scumbag (account holder: user6)

In Figure 11 we show the distribution of account holders involved in multiple instances of attacking another person. We observe that there are a significant number of account holders ($\sim 13\%$) who attack more than once and if we observe the distribution among them, while a majority of them attack 2-5 times, there are also account holders who attack more than 10 times. This scenario of multiple incivil posts seems alarming as it might be one of the prime reason for users to quit social media sites. Thus, it is important to detect these account holders' accounts and take necessary preventive measures.

We also observe that the same target is attacked multiple times. In Figure 12, we see that there are a significant number of targets ($\sim 19\%$) who get attacked multiple times and a majority of them experience 2-3 times of attack. However, there are also a small fraction of targets who are attacked more than 10 times.

7.2 Reputation scores

Reputation score is a measure that determines the reputation of a profile in social media and is defined as $\frac{\#followers}{\#followers+\#friends}$. A celebrity who tends to have a very large followership but follows less people, i.e., having $followers \gg friends$ will have a reputation score close to 1. Figure 13a shows the average reputation scores for the account holders, the targets, the multiple-time insulting account holders and the multiple-time targets. The plot suggests that targets who are attacked multiple times have the highest reputation scores and are therefore celebrities with high probability.



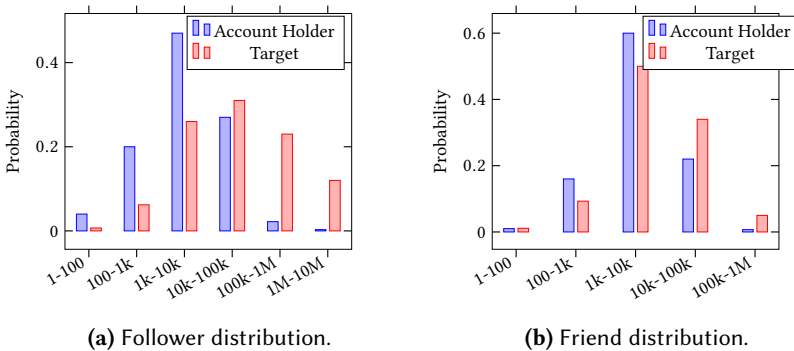
(a) Reputation score comparison between account holders, targets, account holders offending multiple times and multiple-time targets. (b) Reputation score comparison for account holder-target pairs. (c) Reputation score comparison for account holder-target pairs appearing multiple times and pairs appearing once.

Fig. 13. Reputation score comparison

Moreover, the targets in general tend to have higher reputation scores than the account holders. We next consider the account holder-target pairs. Figure 13b shows the distribution of the reputation score ratio $\left(\frac{repscore(target)}{repscore(accountholder)}\right)$ for the account holder-target pairs. It is quite evident from the result that there are more number of account holder-target pairs with reputation score ratio more than 1. Therefore, in general, the targets are more reputed than account holders. Figure 13c shows the distribution of reputation score ratios for the account holder-target pairs appearing multiple times and pairs appearing once. The pairs occurring multiple times tend to have greater reputation score ratio with higher probability. Thus, targets who are attacked by the same account holder again and again seem to have higher number of followers, and are highly likely to be celebrities.

7.3 Follower/followee properties

We further study the follower/followee behavior of multiple-time insulting account holders and targets. Figure 14a shows a comparison between the followers of account holders who attack multiple times and targets who are targeted multiple times. The graph shows that the targets have higher number of followers in general as compared to the account holders (p -value of significance is $< 10^{-5}$). Many of the targets' follower count falls in the 1M – 10M range, whereas more than 40% of the account holders' follower count lies in the 1K-10K bucket. Figure 14b compares the friends of multiple-time insulting account holders and targets. It is evident from the distribution that account holders have larger number of friends in general with high probability (p -value of significance is $< 10^{-5}$).



(a) Follower distribution. (b) Friend distribution.

Fig. 14. Distribution of various properties for multiple-time account holders and multiple-time targets.

8 DISCUSSIONS AND CONCLUSIONS

8.1 Implications of our work

There are several implications of this study. We report these in the following.

Mitigating the spread of negativity: Incivility has high impact on the spread of negativity among communities in social media platforms. In prior research works [23, 24, 101], it has been shown that negative behavior can persist and spread across a community (if not controlled) due to its reinforcing nature. Opinion disagreements can act as early indicators for stopping the spread of such negativities. Our proposed framework can be efficiently used to detect opinion disagreements that might lead to incivilities early in time with minimal features. Since lot of conversations around controversial topics or policy issues, e.g., global warming, vaccination, abortion, immigration etc. on social media generally lead to opinion disagreements, our framework can be tuned to detect the level of toxicity or incivility in the posts so that only the ones which are below a threshold (can be a system parameter) are detected. This will increase the system efficiency of detecting potential incivil incidents. Early detection of these opinion disagreements can act as alerts that can be sent to content moderators (bots or humans) so that they monitor the conflicts and take necessary actions to mitigate the further spread of negativity in the community.

Social and policy implications: Our work also has important implications to law, public policy and the society. Wolfson in context of American concept of free speech and political scenario, has argued in his book [103] that it is almost impossible to separate the good speech or the bad speech/incivility/abusiveness. According to the First Amendment to the United States Constitution, one has to provide adequate opportunities to express differing opinions and engage in public political debate. However, Wolfson also notes that in case of private individuals, these differing opinions impact emotional health of the targeted individual and hence must be prohibited. In contrast, incivility directed toward a group or community has the potential to mobilize a larger number of individuals and can have devastating consequences. These open up the debate for the need to censorship, regulations or Government laws.

Design of online platforms to reduce conflicts: It might not be always appropriate to attribute the strong sentiment expressed by a target to be representative of his/her internal characteristics; in contrast, it might be because of a very specific and one-time external factor, for instance a “very bad day” of the target for which he/she might hold the named entity responsible eventually causing him/her to post a (one-off) tweet expressing a strong sentiment toward that named entity. Our work sheds light on how we might design discussion platforms that minimize the number of opinion conflicts due to such incidents.

Early detection of opinion conflict might be used by the platform moderators to alter the context of the discussion by possibly selectively hiding strong sentiments toward entities and prioritizing constructive comments. This may directly enhance civility in the platform making account holders less likely to be invective (similar observations in the context of trolling and online harassment has been made in earlier works [8, 23]).

In addition, community norms might be introduced to reduce conflicts. For instance, the moment someone is found to express strong sentiments toward a certain entity, the platform may raise alerts/reminders of ethical standards or cite past moral actions.

We believe that many of these efforts can come directly from the researchers and designers in the CSCW community so as to make the online experience of user more safe and accommodating.

8.2 Conclusions

In this paper, we study the behavioral aspects of the targets and account holders and try to understand various factors associated with incivility incidents. We then automate the incivility detection process by developing a deep learning model with character level LSTM and CNNs and incorporating information related to entity-specific opinion conflicts. Our model achieves an accuracy of 93.3% with an F1-score of 0.82 which significantly outperforms the best performing baseline models achieving 4.9%, 6.5% improvements w.r.t accuracy, F1-score respectively.

Insights: We gained several insights from our analysis and automation. Some of these are -

- (i) Opinion conflicts among the target and the account holder at early stages can lead to eventual instances of incivility;
- (ii) Targets are usually more popular with higher reputation compared to account holders; this is opposite to instances of cyberbullying where the perpetrator is usually more reputed and powerful;
- (iii) Account holders tend to post tweets that are inflicted with negative sentiments/emotions, “swear”, “sex”, “religion” and “death” words;
- (iv) Targets usually post tweets rich in positive emotions;
- (v) There are extensive evidences of repeated incivility as well as repeated targeting, sometimes each of these up to 10 times. This could be very alarming for the social media sites and suitable automatic procedures should be built in to contain such cases early in time. Our deep learning model could be a first step toward this enterprise.

8.3 Future works

There are quite a few interesting directions that can be explored in future. One such direction could be to look into the problem of incivility from perspectives of various *events* and *topics* and study the effect of such topics/events on the incivility dynamics. Another interesting direction could be to study the temporal characteristics of incivility incidents on social media and identify the factors leading to rise in incivility incidents [6, 60, 84]. Apart from understanding the cause of incivility and thereby detecting it, one can also develop alert-systems that can subsequently take action to mitigate the impact of opinion disagreements leading to incivilities.

REFERENCES

- [1] Aseel Addawood, Rezvaneh Rezapour, Omid Abdar, and Jana Diesner. 2017. Telling Apart Tweets Associated with Controversial versus Non-Controversial Topics. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 32–41.
- [2] Robert H Aseltine Jr, Susan Gore, and Jennifer Gordon. 2000. Life stress, anger and anxiety, and delinquency: An empirical test of general strain theory. *Journal of Health and Social Behavior* (2000), 256–275.
- [3] Paul Baker. 2001. Moral panic and alternative identity construction in Usenet. *Journal of Computer-Mediated Communication* 7, 1 (2001), JCMC711.
- [4] Kaspar Beelen, Evangelos Kanoulas, and Bob van de Velde. 2017. Detecting Controversies in Online News Media (*SIGIR '17*). 1069–1072.
- [5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.
- [6] Annabella Biancheri. 2017. Cyberbullying on the rise; social media to blame. <http://lobbyobserver.org/lobbyo/2016/11/cyberbullying-on-the-rise-due-to-recent-technological-advances/>. (2017). [Online].
- [7] Amy Binns. 2012. DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities. *Journalism Practice* 6, 4 (2012), 547–562.
- [8] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (2017), 24:1–24:19 pages.
- [9] BlogXilla. 2013. 8 Of The Most Cyberbullied Celebrities On The Web (LIST). <https://globalgrind.cassiuslife.com/1969918/8-most-cyberbullied-celebrities-web-social-media-list/>. (2013). [Online].

- [10] Bouke Bommerson. 2015. Machine learning to classify bullying messages on twitter. . https://www.authorea.com/users/40545/articles/46776/_show_article. (2015).
- [11] Danah Boyd. 2014. *It's complicated: The social lives of networked teens*. Yale University Press.
- [12] Lisa Broidy and Robert Agnew. 1997. Gender and crime: A general strain theory perspective. *Journal of research in crime and delinquency* 34, 3 (1997), 275–306.
- [13] Elia Bruni, N Tram, Marco Baroni, et al. 2014. Multimodal distributional semantics. *The Journal of Artificial Intelligence Research* 49 (2014), 1–47.
- [14] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and Individual Differences* 67 (2014), 97–102.
- [15] Amanda Burgess-Proctor, Justin W Patchin, and Sameer Hinduja. 2009. Cyberbullying and online harassment: Reconceptualizing the victimization of adolescent girls. *Female crime victims: Reality reconsidered* (2009), 153–175.
- [16] Marilyn A Campbell. 2005. Cyber Bullying: An Old Problem in a New Guise?. *Australian journal of Guidance and Counselling* 15, 01 (2005), 68–76.
- [17] Kellie E Carlyle and Kenneth J Steinman. 2007. Demographic Differences in the Prevalence, Co-Occurrence, and Correlates of Adolescent Bullying at School. *Journal of School Health* 77, 9 (2007), 623–629.
- [18] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Hate is not Binary: Studying Abusive Behavior of# GamerGate on Twitter. *arXiv preprint arXiv:1705.03345* (2017).
- [19] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *WebSci*. 13 – 22.
- [20] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring# GamerGate: A Tale of Hate, Sexism, and Bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1285–1290.
- [21] Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Presenting a labelled dataset for real-time detection of abusive user posts. In *Proceedings of the International Conference on Web Intelligence*. ACM, 884–890.
- [22] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proc. of PASSAT*. 71–80.
- [23] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *CSCW*, Vol. 2017. 1217.
- [24] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *ICWSM*.
- [25] June F Chisholm. 2006. Cyberspace violence against girls and adolescent females. *Annals of the New York Academy of Sciences* 1087, 1 (2006), 74–89.
- [26] Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55 (2004), 591–621.
- [27] Dan Claredon. 2017. 11 Celebrities Cyberbullied Off Social Media: Ed Sheeran, Normani Kordei, & More. <http://www.wetpaint.com/cyberbullied-celebrities-twitter-1507798/>. (2017). [Online].
- [28] Lucie Corcoran, Conor Mc Guckin, and Garry Prentice. 2015. Cyberbullying or Cyber Aggression?: A Review of Existing Definitions of Cyber-Based Peer-to-Peer Aggression. *Societies* 5, 2 (2015), 245–255.
- [29] Maral Dadvar, Franciska MG de Jong, RJF Ordeman, and RB Trieschnigg. 2012. Improved cyberbullying detection using gender information. (2012).
- [30] Noelle Devoe. 2016. 8 Celebrities Who Have Quit Social Media. <https://www.seventeen.com/celebrity/news/a42328/celebrities-who-have-quit-social-media/>. (2016). [Online].
- [31] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *TIIS* 2, 3 (2012), 18.
- [32] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011).
- [33] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. ACM, 29–30.
- [34] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification.. In *Proc. of ACL*. 49–54.
- [35] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. In *ICWSM*.
- [36] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In *ICWSM*.

- [37] Nels Ericson. 2001. *Addressing the problem of juvenile bullying*. US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention Washington, DC.
- [38] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of ACL (ACL '05)*. 363–370.
- [39] David Finkelhor, Kimberly J Mitchell, and Janis Wolak. 2000. Online Victimization: A Report on the Nation's Youth. (2000).
- [40] Sharon Florentine. 2016. Twitter needs to stop the harassment. <https://www.cio.com/article/3098441/leadership-management/twitter-needs-to-stop-the-harassment-leslie-jones.html>. (2016). [Online].
- [41] Rolf Fredheim, Alfred Moore, and John Naughton. 2015. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference (WebSci '15)*.
- [42] Dorothy Wunmi Grigg. 2010. Cyber-aggression: Definition and concept of cyberbullying. *J. of psychologists and counsellors in schools* 20, 2 (2010), 143–156.
- [43] Joshua Guberman, Carol Schmitz, and Libby Hemphill. 2016. Quantifying toxicity and verbal violence on Twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, 277–280.
- [44] Umair Haque. 2016. The Reason Twitter's Losing Active Users. <https://hbr.org/2016/02/the-reason-twitters-losing-active-users>. (2016). [Online].
- [45] Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. (2010).
- [46] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *The information society* 18, 5 (2002), 371–384.
- [47] S Hinduja and JW Patchin. 2012. School climate 2.0: Reducing teen technology misuse by reshaping the environment. (2012).
- [48] Sameer Hinduja and Justin W Patchin. 2007. Offline consequences of online victimization: School violence and delinquency. *Journal of school violence* 6, 3 (2007), 89–112.
- [49] Sameer Hinduja and Justin W Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant behavior* 29, 2 (2008), 129–156.
- [50] Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research* 14, 3 (2010), 206–221.
- [51] Sameer Hinduja and Justin W Patchin. 2015. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying (2nd edition)*. Thousand Oaks, CA: Sage Publications.
- [52] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [53] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv preprint arXiv:1702.08138* (2017).
- [54] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. 2014. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *ASONAM '14*. 244–252.
- [55] Homa Hosseinmardi, Shaosong Li, Zhili Yang, Qin Lv, Rahat Ibn Rafiq, Richard Han, and Shivakant Mishra. 2014. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. 355–362.
- [56] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishr. 2015. Prediction of Cyberbullying Incidents on the Instagram Social Network. *arXiv preprint arXiv:1508.06257* (2015).
- [57] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).
- [58] John W Jones and G Anne Bogat. 1978. Air pollution and human aggression. *Psychological Reports* 43, 3 (1978), 721–722.
- [59] Jaana Juvonen and Elisheva F Gross. 2008. Extending the school grounds?—Bullying experiences in cyberspace. *Journal of School health* 78, 9 (2008), 496–505.
- [60] Maria Konnikova. 2015. How the Internet Has Changed Bullying. <https://www.newyorker.com/science/maria-konnikova/how-the-internet-has-changed-bullying>. (2015). [Online; The New Yorker].
- [61] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proc. of ACM web science conference*. ACM, 195–204.
- [62] Robin M Kowalski and Susan P Limber. 2007. Electronic bullying among middle school students. *Journal of adolescent health* 41, 6 (2007), S22–S30.
- [63] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. 2012. *Cyberbullying: Bullying in the digital age*. John Wiley & Sons.
- [64] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing*

Systems (CHI '15). 3739–3748.

- [65] Amanda Lenhart. 2007. Cyberbullying and online teens. (2007).
- [66] Qing Li. 2007. New bottle but old wine: A research of cyberbullying in schools. *Computers in human behavior* 23, 4 (2007), 1777–1791.
- [67] Alice E Marwick et al. 2011. The drama! Teen conflict, gossip, and bullying in networked publics. (2011).
- [68] Adrienne Massanari. 2017. #Gamergate and The Fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346.
- [69] Paul Mazerolle, Velmer S Burton, Francis T Cullen, T David Evans, and Gary L Payne. 2000. Strain, anger, and delinquent adaptations specifying general strain theory. *Journal of Criminal Justice* 28, 2 (2000), 89–101.
- [70] Paul Mazerolle and Alex Piquero. 1998. Linking exposure to strain with anger: An investigation of deviant adaptations. *Journal of Criminal Justice* 26, 3 (1998), 195–211.
- [71] David McNamee. 2016. Cyberbullying 'causes suicidal thoughts in kids more than traditional bullying'. <https://www.medicalnewstoday.com/articles/273788.php>. (2016). [Online].
- [72] Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More Than Words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 299–303.
- [73] Ersilia Menesini and Annalaura Nocentini. 2009. Cyberbullying definition and measurement: Some critical considerations. *Journal of Psychology* 217, 4 (2009), 230–232.
- [74] Faye Mishna, Charlene Cook, Tahany Gadalla, Joanne Daciuk, and Steven Solomon. 2010. Cyber bullying behaviors among middle and high school students. *American Journal of Orthopsychiatry* 80, 3 (2010), 362–374.
- [75] Vinita Nahar, Sayan Unankard, Xue Li, and Chaoyi Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Asia-Pacific Web Conference*. 767–774.
- [76] Tonja R Nansel, Mary Overpeck, Ramani S Pilla, W June Ruan, Bruce Simons-Morton, and Peter Scheidt. 2001. Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *Jama* 285, 16 (2001), 2094–2100.
- [77] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.
- [78] Justin W Patchin and Sameer Hinduja. 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice* 4, 2 (2006), 148–169.
- [79] Justin W Patchin and Sameer Hinduja. 2011. Traditional and nontraditional bullying among youth: A test of general strain theory. *Youth & Society* 43, 2 (2011), 727–751.
- [80] Justin W Patchin and Sameer Hinduja. 2012. An update and synthesis of the research. *Cyberbullying prevention and response: Expert perspectives* (2012), 13.
- [81] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1125–1135.
- [82] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.
- [83] Quang Anh Phan and Vanessa Tan. 2017. Play with Bad Words: A Content Analysis of Profanity in Video Games. *Acta Ludica-International Journal of Game Studies* 1, 1 (2017), 7–30.
- [84] Team Rawhide. 2017. Teen Cyberbullying and Social Media Use on the Rise [INFOGRAPHIC]. <http://www.rawhide.org/blog/wellness/teen-cyberbullying-and-social-media-use-on-the-rise/>. (2017). [Online].
- [85] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *ICMLA*, Vol. 2. 241–244.
- [86] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP*.
- [87] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open Domain Event Extraction from Twitter. In *KDD*.
- [88] Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. 2017. Detecting Nastiness in Social Media. In *Proceedings of the First Workshop on Abusive Language Online*. 63–72.
- [89] Huascar Sanchez and Shreyas Kumar. 2011. Twitter bullying detection. *ser. NSDI* 12 (2011), 15–15.
- [90] Shari Kessel Schneider, Lydia O'Donnell, Ann Stueve, and Robert WS Coulter. 2012. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American Journal of Public Health* 102, 1 (2012), 171–177.
- [91] Dorothy Seals and Jerry Young. 2003. Bullying and victimization: Prevalence and relationship to gender, grade level, ethnicity, self-esteem, and depression. *Adolescence* 38, 152 (2003), 735.
- [92] Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science* 36, 3 (2010), 357–370.
- [93] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media.. In *ICWSM*. 687–690.

- [94] Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. " They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools.. In *CSCW*. 1210–1216.
- [95] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry* 49, 4 (2008), 376–385.
- [96] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective LSTMs for Target-Dependent Sentiment Classification. *arXiv preprint arXiv:1512.01100* (2015).
- [97] Robert S Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior* 26, 3 (2010), 277–287.
- [98] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics (HUSO 2015)*. IARIA, 13–18.
- [99] Heidi Vandebosch and Katrien Van Cleemput. 2008. Defining cyberbullying: A qualitative research into the perceptions of youngsters. *CyberPsychology & Behavior* 11, 4 (2008), 499–503.
- [100] Jing Wang, Ronald J Iannotti, and Tonja R Nansel. 2009. School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. *Journal of Adolescent health* 45, 4 (2009), 368–375.
- [101] Robb Willer, Ko Kuwabara, and Michael W Macy. 2009. The false enforcement of unpopular norms. *Amer. J. Sociology* 115, 2 (2009), 451–490.
- [102] Kirk R Williams and Nancy G Guerra. 2007. Prevalence and predictors of internet bullying. *Journal of adolescent health* 41, 6 (2007), S14–S21.
- [103] Nicholas Wolfson. 1997. Hate Speech, Sex Speech, Free Speech. (1997).
- [104] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proc. of NAACL*. 656–666.
- [105] Michele L Ybarra, Danah Boyd, Josephine D Korchmaros, and Jay Koby Oppenheim. 2012. Defining and measuring cyberbullying within the larger context of bullying victimization. *Journal of Adolescent Health* 51, 1 (2012), 53–58.
- [106] Michele L Ybarra and Kimberly J Mitchell. 2004. Youth engaging in online harassment: Associations with caregiver-child relationships, Internet use, and personal characteristics. *Journal of adolescence* 27, 3 (2004), 319–336.
- [107] Michele L Ybarra and Kimberly J Mitchell. 2007. Prevalence and frequency of Internet harassment instigation: Implications for adolescent health. *Journal of Adolescent Health* 41, 2 (2007), 189–195.
- [108] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proc. of the Content Analysis in the WEB 2* (2009), 1–7.
- [109] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*. 649–657.
- [110] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. (2016).

Received April 2018; revised July 2018; accepted September 2018